MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2015/2016

TMA 7021 – DATA MINING AND ANALYTICS

(All sections / Groups)

21 SEPTEMBER 2015
8.00 p.m – 10.00 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of **6** pages with four questions only.
2. Answer **ALL** questions.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

(a) Differentiate between *descriptive* data mining and *predictive* data mining. List at least one technique for *descriptive* data mining and *predictive* data mining.

[2 marks]

(b) Apply binning method to smooth for the following data. The depth of each bin must be FOUR (4).

22	52	23	44	35	62	12	40	58	61	62	70

i) Bin means

[2 marks]

ii) Bin boundaries

[2 marks]

(c) Study the following dataset carefully and write your answer in **R** codes.

Data Frame name = **`dt_Student`**

Name	**Age**	**CGPA**
Ting	20	3.5
Chong	NA	4.4
David	25	NA
Elyne	18	0.5
Fatimah	NA	NA

* *missing value is represented as "NA"*

(i) Subset all records that have missing values for Age.

[2 marks]

(ii) Replace all the missing values for CGPA with the mean of CGPA.

[2 marks]

Continued...

QUESTION 2

(a) Study the table below.

Transaction	Book Type	Fruit Type
1	1,3,4	O, T
2	2,3	O, T
3	3	T
4	1,3,5,6	O
5	2,4	T
6	1,3	O
7	2,3,5	T
8	1,2	O

Calculate the *support* and *confidence* of the following association rules:

[2 marks]

```
purchase(X, "Book 1") => purchase (X, "Book 3")
purchase(X, "Book 3") => purchase (X, "Fruit O")
```

(b) Below is the output of association rule mining using R. Discuss your conclusion based on the given output.

```
    lhs                rhs             support confidence  lift
1   {Class=2nd,
     Age=Child}    => {Survived=Yes}    0.011      1.000 3.096
2   {Class=2nd,
     Sex=Female,
     Age=Child}    => {Survived=Yes}    0.006      1.000 3.096
```

[2 marks]

(c) Differentiate *K-Nearest Neighbours* and *K-Means*. Using the following dataset, use *K-Means* to find two clusters from this set of data. Start with centroids {1} and {390}. You must show the steps in each iteration.

```
{2, 4, 5, 9, 13, 28, 34, 50, 70, 120, 300, 389}
```

[6 marks]

Continued...

QUESTION 3

(a) The term-frequency table below shows frequency of terms $T_1,\ldots,T_5$ appears in website $W_1,\ldots,W_4$.

	W_1	W_2	W_3	W_4
T_1	20	12	40	20
T_2	15	2	20	5
T_3	30	8	5	15
T_4	5	15	4	55
T_5	16	5	10	2

By using *Cosine* measure, determine which term ($T_1,\ldots,T_5$) is *most* related to T_1?

[4 marks]

(b) Define outlier and discuss the different approaches for outlier analysis.

[2 marks]

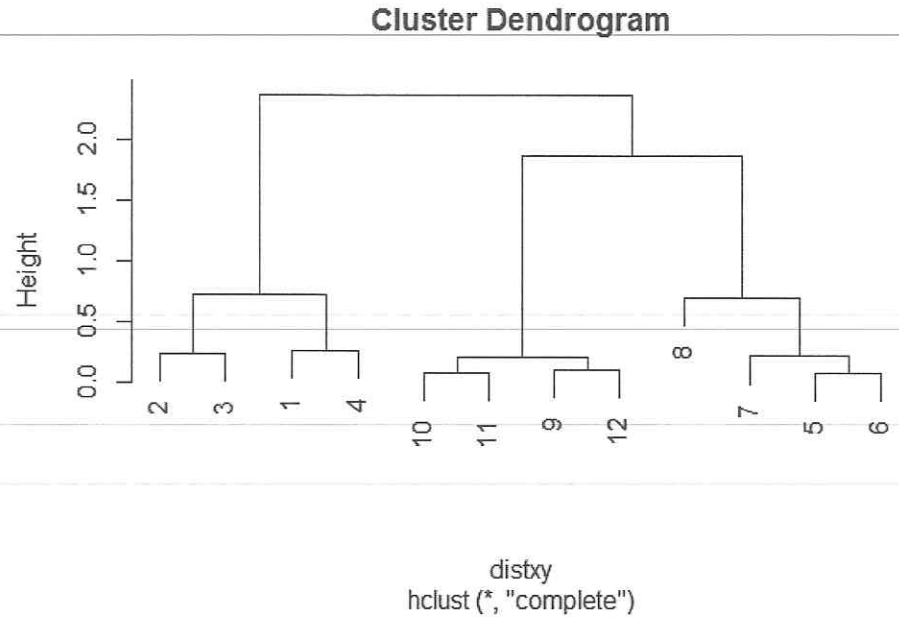
(c) The information below presents the summary of "Income" of a particular company. Discuss why anomaly exists in the information.

```
> summary(Income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -8.7    14.6    35.0    53.5    67.0   615.0
```

[2 marks]

Continued...

(d) Study the following cluster dendrogram for hierarchical clustering.



[2 marks]

What will happen if the following R script is executed?

```
rect.hclust(hClusters, k=2, border="red")
```

Continued...

QUESTION 4

(d) Define and discuss the three properties of *dirty data*.

[3 marks]

(a) Study the following information carefully.

Measuring the popularity of a social media mobile application is very difficult. Many factors needs to be considered. Following is a list of variables and evidence as input a Bayesian Network to predict the popularity of that application.

Variables that influence popularity of social media mobile application:
Trust, *performance, reputation.*

Evidence for Trust:
Number of tweets, *number of Followers*, *number of Mention*, *number of Retweets.*

Assuming that there is *no dependency between the variables*, and *no dependency between the evidence*. Create the Bayesian Network using the above information.

[5 marks]

(b) Name the possible states for all the *non-observable* nodes in the network that you have created in Question 4(b).

[2 marks]

End of Pages.